AMENDMENTS TO THE CLAIMS

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1. (Currently amended) A computer implemented system that facilitates building a statistical model for a computer readable data set, comprising:

a first training ~~algorithm~~ method that efficiently builds a rough statistical model from a subset of the computer readable data set capable of statistical characterization;

an evaluation component that evaluates the rough statistical model to determine whether the subset of the computer readable data set is an appropriate subset to be utilized to build a refined statistical model for the computer readable data set based at least in part on stopping criterion to facilitate reducing cost of clustering data relative to the computer readable data set;

a second training ~~algorithm~~ method that builds [[a]] the refined statistical model for the computer readable data set from the subset if the subset is deemed appropriate by the evaluation component, the refined statistical model provides a more accurate modeling of the subset than the rough statistical model and facilitates determining ~~discovers~~ good clustering of data for a fixed number of clusters based at least in part on predefined accuracy criteria to facilitate clustering of data relative to the computer readable data set, wherein the clustered data is provided; and

a data scheduler that, based at least in part on a data policy, adaptively controls the size of subsets for which the first training ~~algorithm~~ method is applied to facilitate building ~~an accurate~~ the refined statistical model.

2. (Canceled)

3. (Currently amended) The system of claim 1, the data scheduler increases the size of the subset to provide a larger aggregate subset of the data set if the rough statistical model is unacceptable, the first training ~~algorithm~~ method efficiently builds the rough statistical model for each larger aggregate subset of the data until the evaluation component determines the resulting rough statistical model to be acceptable.

4. (Currently amended) The system of claim 3, the acceptability of each rough statistical model is determined based at least in part on a stopping criterion functionally related to an expected incremental benefit and a cost associated with increasing the size of the aggregate subset of the data set.

5. (Currently amended) The system of claim 4, the cost of the stopping criterion is functionally related to at least one of time associated with evaluating an aggregate data subset of increased size [[and]] or size of the aggregated subset of the data.

6. (Currently amended) The system of claim 4, the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

c_1, c_2, and c_3 are constants determined based on application of the second training ~~algorithm~~ method relative to a first subset of the data set,

I_1 is a number of iterations for the second training ~~algorithm~~ method, when applied to the first subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is the number of iterations for the first training ~~algorithm~~ method when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$,

λ is a user determined stopping threshold.

7. (Currently amended) The system of claim 4, the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))+\delta-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

δ is an offset associated with a difference in log likelihood for holdout data when evaluated for models built on a first subset of the training data set by the respective first and second training ~~algorithms~~ methods,

c_1, c_2, and c_3 are constants determined based at least in part on application of the second training ~~algorithm~~ method relative to a first subset of the data set,

I_1 is a number of iterations for the second training ~~algorithm~~ method, when applied to the first subset,

$\bar{J}_n=\frac{1}{n}\sum_{i=1}^{n}J_i$, and

J_i is the number of iterations for the first training ~~algorithm~~ method when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

8. (Currently amended) The system of claim 1, the first training ~~algorithm~~ method further comprises an iterative ~~algorithm~~ method, which builds the rough statistical model for the subset of the data set according to an associated training policy.

9. (Currently amended) The system of claim 8, the first training ~~algorithm~~ method further comprises an associated training policy that defines parameter initialization of the first training ~~algorithm~~ method for each subset of the data set.

10. (Currently amended) The system of claim 9, the training policy associated with the first training ~~algorithm~~ method further controls parameter initialization of the first training ~~algorithm~~ method, such that at least some of the parameters computed for a previous subset of the data are employed to initialize the first training ~~algorithm~~ method for a subsequent larger aggregate subset of the data.

11. (Currently amended) The system of claim 9, the first training ~~algorithm~~ method is initialized by the same parameter values for each subset of the data subset.

12. (Currently amended) The system of claim 9, the training policy sets the iterative ~~algorithm~~ method to perform a fixed number of at least one iteration.

13. (Currently amended) The system of claim 12, the training policy sets the iterative ~~algorithm~~ method to perform a single iteration.

14. (Currently amended) The system of claim 12, the second training ~~algorithm~~ method further comprises an iterative ~~algorithm~~ method that operates according to an associated training policy, so as to produce a more accurate statistical model for the appropriate subset of the data set than the first training ~~algorithm~~ method.

15. (Currently amended) The system of claim 14, the iterative ~~algorithm~~ method associated with at least one of the first [[and]] or second training ~~algorithms~~ methods is an Expectation and Maximization ~~algorithm~~ method.

16. (Currently amended) The system of claim 8, the training policy associated with the iterative ~~algorithm~~ method of the first training ~~algorithm~~ method controls the iterative ~~algorithm~~ method to run until an associated convergence criterion is satisfied.

17. (Currently amended) The system of claim 16, second training ~~algorithm~~ method further comprises an iterative ~~algorithm~~ method, which builds the refined statistical model for the appropriate subset of the data set according to an associated training policy.

18. (Currently amended) The system of claim 17, the training policy associated with the iterative ~~algorithm~~ method of the second training ~~algorithm~~ method controls the respective iterative ~~algorithm~~ method to run until an associated convergence criterion is satisfied, the convergence criterion associated with the second training ~~algorithm~~ method provides improved model quality relative to the convergence criterion associated with the first training ~~algorithm~~ method.

19. (Currently amended) A computer implemented system programmed to facilitate building a statistical model, comprising:

a first parameter estimation ~~algorithm~~ protocol that efficiently builds a rough statistical model from a subset of a computer readable data set based at least in part on a training policy associated therewith, the computer readable data set is statistically characterizable; [[and]]

an evaluation component that determines whether the subset of data from which the rough statistical model was built is an acceptable size for building the statistical model to characterize the data set, the evaluation component utilizes a stopping criterion that is functionally related to an expected incremental benefit and an expected incremental cost associated with increasing the size of the subset of data to facilitate determining whether the rough statistical model is an acceptable size and to facilitate reducing cost of clustering data relative to the computer readable data set; and

a second parameter estimation ~~algorithm~~ protocol that builds a refined statistical model for the data set from the subset if determined to have the acceptable size, the second parameter estimation ~~algorithm~~ protocol having an associated training policy, which enables the second parameter estimation ~~algorithm~~ protocol to build the refined statistical model to be a more accurate statistical model than the first parameter estimation ~~algorithm~~ protocol, the ~~more accurate~~ refined statistical model employed to identify clusters of data within the computer readable data set to facilitate clustering data relative to the computer readable data set, wherein the clustered data is provided.

20. (Currently amended) The system of claim 19, further comprising a data scheduler that increases the size of the subset of the data set to provide a larger aggregate subset of the data set if the rough statistical model is unacceptable, the first parameter estimation ~~algorithm~~ protocol efficiently builds a rough statistical model for each larger aggregate subset until a resulting rough statistical model built therefrom is determined to be acceptable.

21. (Currently amended) The system of claim 19, the first parameter estimation ~~algorithm~~ protocol further comprises an iterative ~~algorithm~~ protocol that builds the rough statistical model for each subset of the data set according to the associated training policy.

22. (Currently amended) The system of claim 21, the training policy for the first parameter estimation ~~algorithm~~ protocol is operative to control parameter initialization for the first parameter estimation ~~algorithm~~ protocol, such that at least some of the parameters computed for a previous subset of the data are employed to initialize the first parameter estimation ~~algorithm~~ protocol for a subsequent larger aggregate subset of the data set.

23. (Currently amended) The system of claim 21, the first parameter estimation ~~algorithm~~ protocol is initialized by the same parameter values for each subset of the data subset.

24. (Currently amended) The system of claim 21, the training policy associated with first parameter estimation ~~algorithm~~ protocol controls the iterative ~~algorithm~~ protocol of the first parameter estimation ~~algorithm~~ protocol to perform a fixed number of at least one iteration, the second training ~~algorithm~~ protocol further comprising an iterative ~~algorithm~~ protocol, which is operative to perform a greater number of iterations than the iterative ~~algorithm~~ protocol of the first training ~~algorithm~~ protocol based at least in part on a training policy associated with the second parameter estimation ~~algorithm~~ protocol.

25. (Currently amended) The system of claim 21, the training policy associated with the iterative ~~algorithm~~ protocol of the first parameter estimation ~~algorithm~~ protocol controls the iterative ~~algorithm~~ protocol to run until an associated convergence threshold is satisfied, the second training ~~algorithm~~ protocol further comprises an iterative ~~algorithm~~ protocol, the training policy associated with the iterative ~~algorithm~~ protocol of the second parameter estimation ~~algorithm~~ protocol being operative to control the respective iterative ~~algorithm~~ protocol to run until an associated convergence threshold is satisfied, the convergence threshold associated with the second parameter estimation ~~algorithm~~ protocol is less than the convergence threshold associated with the first parameter estimation ~~algorithm~~ protocol.

26. (Canceled)

27. (Currently amended) The system of claim 19, the cost of the stopping criterion is functionally related to at least one of time associated with evaluating the model for a larger subset of data [[and]] or size of the larger subset of the data.

28. (Currently amended) The system of claim 19, the stopping criterion is defined by

$$\left(\frac{l(D_{HO} \mid \theta(D_n)) - l(D_{HO} \mid \theta(D_{n-1}))}{l(D_{HO} \mid \theta(D_n)) - l(D_{HO} \mid \theta_{BASE}(D_n))}\right) \frac{1}{c_1(I_1 - \bar{J}_n) \mid \Delta D_{n+1} \mid + c_2(I_1 - \bar{J}_n) + c_1\bar{J}_n \mid D_{n+1} \mid + c_2\bar{J}_n + c_3} < \lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ protocol on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ protocol on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

c_1, c_2, and c_3 are constants determined based at least in part on application of the second parameter estimation ~~algorithm~~ protocol relative to a first subset of the data set,

I_1 is a number of iterations for the second parameter estimation ~~algorithm~~ protocol, when applied to the first subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is the number of iterations for the first parameter estimation ~~algorithm~~ protocol when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

29. (Currently amended) The system of claim 19, the stopping criterion is defined by

$$\left(\frac{l(D_{HO}\mid\theta(D_n))-l(D_{HO}\mid\theta(D_{n-1}))}{l(D_{HO}\mid\theta(D_n))+\delta-l(D_{HO}\mid\theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1-\bar{J}_n)\mid\Delta D_{n+1}\mid+c_2(I_1-\bar{J}_n)+c_1\bar{J}_n\mid D_{n+1}\mid+c_2\bar{J}_n+c_3}<\lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ protocol on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ protocol on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

δ is an offset associated with a difference in log likelihood for holdout data when evaluated for models built on a first subset of the training data set by the respective first and second training ~~algorithms~~ protocols,

c_1, c_2, and c_3 are constants determined based at least in part on application of the second parameter estimation ~~algorithm~~ protocol relative to a first data subset of the data set,

I_1 is a number of iterations for the second parameter estimation ~~algorithm~~ protocol, when applied to a first data subset,

$\bar{J}_n=\frac{1}{n}\sum_{i=1}^{n}J_i$, and

J_i is the number of iterations for the first parameter estimation ~~algorithm~~ protocol when applied to a data subset D_i,

$|D_{n+1}|$ is the size of data set D_{n+1},

$|\Delta D_{n+1}|$ is the increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

30. (Currently amended) A computer implemented learning curve method to facilitate building a statistical model, comprising:

choosing a subset of a computer readable data set that can be characterized statistically;

employing a first training ~~algorithm~~ method to build a rough statistical model to characterize the subset;

evaluating the rough statistical model for acceptability;

if the rough statistical model is unacceptable, repeatedly increasing the size of the subset of data to provide an aggregate data set, building another rough statistical model to characterize the aggregate subset, and reevaluating the other rough statistical model, the acceptability of each rough statistical model based at least in part on a stopping criterion functionally related to an expected incremental benefit and an expected incremental cost associated with increasing the size of the aggregate subset in order to facilitate reducing cost associated with clustering data relative to the computer readable data set; and

if the rough statistical model is acceptable, employing a second training ~~algorithm~~ method to build a refined statistical model based at least in part on the aggregate data set, the second training ~~algorithm~~ method being different from the first training ~~algorithm~~ method, the refined statistical model identifies data clusters contained in the computer readable data set to facilitate clustering of data relative to the computer readable data set, wherein the clustered data is provided.

31. (Canceled)

32. (Currently amended) The system of claim 30, the cost of the stopping criterion is functionally related to at least one of time associated with evaluating an aggregate data subset of increased size [[and]] or size of the aggregate subset of the data.

33. (Currently amended) The system of claim 30, the stopping criterion is defined by

$$\left(\frac{l(D_{HO} \mid \theta(D_n)) - l(D_{HO} \mid \theta(D_{n-1}))}{l(D_{HO} \mid \theta(D_n)) - l(D_{HO} \mid \theta_{BASE}(D_n))}\right)\frac{1}{c_1(I_1 - \bar{J}_n) \mid \Delta D_{n+1} \mid + c_2(I_1 - \bar{J}_n) + c_1\bar{J}_n \mid D_{n+1} \mid + c_2\bar{J}_n + c_3} < \lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

c_1, c_2, and c_3 are constants determined based at least in part on application of the second parameter estimation ~~algorithm~~ method relative to a first subset of the data set,

I_1 is a number of iterations for the second parameter estimation ~~algorithm~~ method, when applied to the first subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is a number of iterations for the first parameter estimation ~~algorithm~~ method when applied to a data subset D_i,

$|D_{n+1}|$ is a size of data set D_{n+1},

$|\Delta D_{n+1}|$ is an increment in size $|D_{n+1}| - |D_n|$, and

λ is a user determined stopping threshold.

34. (Currently amended) The system of claim 30, the stopping criterion is defined by

$$\left(\frac{l(D_{HO} \mid \theta(D_n)) - l(D_{HO} \mid \theta(D_{n-1}))}{l(D_{HO} \mid \theta(D_n)) + \delta - l(D_{HO} \mid \theta_{BASE}(D_n))}\right) \frac{1}{c_1(I_1 - \bar{J}_n) \mid \Delta D_{n+1} \mid + c_2(I_1 - \bar{J}_n) + c_1 \bar{J}_n \mid D_{n+1} \mid + c_2 \bar{J}_n + c_3} < \lambda$$

where

$l(D_{HO}|\theta(D_n))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a current subset of the training data set,

$l(D_{HO}|\theta(D_{n-1}))$ is a log likelihood for holdout data evaluated for the model built by the first training ~~algorithm~~ method on a previous subset of the training data set,

$l(D_{HO}|\theta_{base}(D_n))$ is a log likelihood for holdout data evaluated for a base model,

δ is an offset associated with the difference in log likelihood for holdout data when evaluated for models built on a first subset of the training data set by the respective first and second training ~~algorithms~~ methods,

c_1, c_2, and c_3 are constants determined based at least in part on application of the second parameter estimation ~~algorithm~~ method relative to a first data subset of the data set,

I_1 is a number of iterations for the second parameter estimation ~~algorithm~~ method, when applied to a first data subset,

$\bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_i$, and

J_i is a number of iterations for the first parameter estimation ~~algorithm~~ method when applied to a data subset D_i,

$| D_{n+1}|$ is a size of data set D_{n+1},

$| \Delta D_{n+1}|$ is an increment in size $| D_{n+1}| - | D_n|$, and

λ is a user determined stopping threshold.

35. (Currently amended) The method of claim 30, the first training ~~algorithm~~ method is more computationally efficient than the second training ~~algorithm~~ method.

36. (Currently amended) The method of claim 30, each instance of model building repeated until obtaining an acceptable rough statistical model by the first training ~~algorithm~~ method employs more efficient and less accurate model building than model building employed by the second training ~~algorithm~~ method that occurs after obtaining the acceptable rough statistical model.

37. (Currently amended) The method of claim 36, each instance of model building repeated until obtaining an acceptable rough statistical model employs the first training ~~algorithm~~ method as an iterative ~~algorithm~~ method that is run to a first convergence criterion, the second training ~~algorithm~~ method employing an iterative ~~algorithm~~ method that is run to a second convergence criterion, which demands more iterations than the first convergence criterion in order to obtain convergence, so that the refined statistical model is more accurate than the rough statistical model built by the first training ~~algorithm~~ method.

38. (Currently amended) The method of claim 36, each instance of model building repeated until obtaining an acceptable rough statistical model employs an iterative ~~algorithm~~ method having a fixed number of at least one iteration, the second training ~~algorithm~~ method employing an iterative ~~algorithm~~ method having a greater number of iterations than the fixed number.

39. (Currently amended) The method of claim 30, further comprising controlling parameter initialization employed in each instance of building a model for the aggregate data set prior to obtaining an acceptable rough statistical model.

40. (Currently amended) The method of claim 39, further comprising initializing the first training ~~algorithm~~ method by the same parameter values for each subset.

41. (Currently amended) The method of claim 39, the controlling further comprises reusing at least some of the parameters computed from a previous instance of model building to initialize a subsequent instance of model building for a subsequent larger aggregate data set prior to obtaining an acceptable rough statistical model.

42. (Currently amended) A computer-readable medium having computer-executable instructions for:

choosing a subset of a computer readable data set;

building a rough statistical model to characterize the subset based at least in part on an associated training policy;

evaluating the rough statistical model for acceptability;

if the rough statistical model is unacceptable, repeatedly increasing the size of the subset of data to provide an aggregate data set, building a rough statistical model to characterize the aggregate subset based at least in part on an associated training policy, and reevaluating the rough statistical model;

building a refined statistical model for the computer readable data set from the aggregate data set if the rough statistical model is determined to be acceptable based at least in part on an associated training policy that includes determining acceptability based at least in part on an expected incremental benefit relative to an expected incremental cost associated with increasing the size of the aggregate data set in order to facilitate reducing cost associated with clustering data relative to the computer readable data set, the refined statistical model more accurately characterizes the aggregate data set; and

utilizing the refined statistical model to identify identifiable clusters in the computer readable data set to facilitate clustering data relative to the computer readable data set, wherein the clustered data is provided.

43. (Canceled)

44. (Currently amended) A computer implemented method to facilitate constructing a statistical model, comprising:

separating computer readable data on a computer readable medium into holdout data set and training data set;

determining a data subset from the training data set by estimating statistical model parameters according to a first training policy and evaluating the estimated statistical model parameters relative to the holdout data set and repeating the estimation and evaluation of statistical model parameters with a larger subset of the training data set until an acceptable quality of the estimated statistical model is established to facilitate reducing cost associated with characterizing clusters relative to the computer readable data;

controlling parameter initialization employed in each estimation of statistical model parameters repeatedly until an acceptable size for the determined data subset is achieved; and

subsequent to establishing the acceptable quality of the estimated statistical model, using the determined data subset to improve the estimated statistical model parameters by employing a second training policy that is more accurate than the first training policy, the estimated model parameters obtained from employment of the second training policy utilized to characterize at least one cluster within the computer readable data to facilitate clustering data relative to the computer readable data, wherein the clustered data is provided.

45. (Currently amended) The method of claim 44, each estimation of model parameters repeated until the acceptable quality of the estimated model is established further comprises employing an iterative ~~algorithm~~ method that is run until a first convergence criterion is satisfied, the estimation of model parameters using the determined data subset further comprising an iterative ~~algorithm~~ method that is run until a second convergence criterion is satisfied, which is operative to provide a better quality of model than the first convergence criterion.

46. (Currently amended) The system of claim 45, the first convergence criterion causes the associated iterative ~~algorithm~~ method to run until a first convergence threshold is satisfied, the second convergence criterion causes the associated iterative ~~algorithm~~ method to run until a second convergence threshold is satisfied, the second convergence threshold being less than the first convergence threshold.

47. (Currently amended) The method of claim 45, at least one of the iterative ~~algorithm~~ method run to the first convergence criterion [[and]] or the iterative ~~algorithm~~ method run to the second convergence criterion is an Expectation and Maximization ~~algorithm~~ method.

48. (Currently amended) The method of claim 44, each estimation of model parameters repeated until the acceptable quality of the estimated model is established employs an iterative ~~algorithm~~ method having a fixed number of at least one iteration, the estimation of model parameters using the determined data subset further employing an iterative ~~algorithm~~ method having a greater number of iterations than the fixed number.

49. (Canceled)

50. (Previously Presented) The method of claim 44, the controlling further comprises reusing at least some of the parameters computed from a previous estimation of model parameters to initialize a subsequent estimation of model parameters for a next larger subset of the training set.

51. (Currently amended) The method of claim 44, each estimation of model parameters repeated until the acceptable quality of the estimated model is established further comprises initializing the first training ~~algorithm~~ method by the same parameter values.

52. (Currently amended) The method of claim 44, further comprising determining the acceptability of the estimated model based at least in part on an expected incremental benefit relative to a cost associated with increasing the size of the subset of the data set.

53. (Currently amended) A computer-readable medium having computer-executable instructions for:

separating computer readable data into a holdout data set and a training data set, the computer readable data is statistically characterizable;

determining a data subset from the training data set by estimating model parameters and controlling model parameter initialization[[,]] according to a first training policy and evaluating the estimated model parameters relative to the holdout data set and repeating the estimation, initialization, and evaluation of model parameters with a next successively larger subset of the training data set until an acceptable quality of the estimated model is established to facilitate reducing cost associated with clustering data relative to the computer readable data;

subsequent to establishing the acceptable quality of the estimated model, using the determined data subset to improve the estimated model parameters by employing a second training policy that is more accurate than the first training policy; and

utilizing the estimated model parameters determined by utilization of the second training policy to identify a cluster in the computer readable data to facilitate clustering data relative to the computer readable data, wherein the clustered data is provided.

54. (Currently amended) A computer implemented method to facilitate constructing a statistical model, comprising:

separating computer readable data into a holdout data set and a training data set, the computer readable data is statistically characterizable;

iteratively estimating statistical model parameters for a subset of the training data set over a fixed number of iterations and evaluating the estimated statistical model parameters relative to the holdout data set;

repeating the estimation and evaluation of statistical model parameters obtained with successively larger subsets of the training data set until an acceptable model quality is established, acceptable model quality determined based at least in part on an expected incremental benefit relative to an expected incremental detriment associated with an increase in size of each larger training subset of the data set in order to facilitate reducing cost associated with clustering data relative to the computer readable data;

after the acceptable model quality is established, iteratively estimating statistical model parameters for the data subset, which provided the acceptable model quality, until a better quality of model is provided relative to a preceding estimation performed over the fixed number of iterations; and

using the better quality model relative to the computer readable data to identify at least a cluster of data within the computer readable data to facilitate clustering data relative to the computer readable data, wherein the at least a cluster of data is provided.

55. (Currently amended) The method of claim 54, at least one of the iterative estimations employs an Expectation and Maximization ~~algorithm~~ method.

56. (Currently amended) The method of claim 54, the estimation that occurs after the acceptable model quality is established, further comprises employing an iterative ~~algorithm~~ method having a greater number of iterations than the fixed number.

57. (Currently amended) The method of claim 54, the estimation of model parameters after the acceptable model quality has been established further comprises employing an iterative ~~algorithm~~ method that is run until a convergence criterion is satisfied, which is operative to provide a better quality of model with the data subset than a preceding estimation employing the fixed number of iterations.

58. (Original) The method of claim 54, further comprising controlling parameter initialization for each estimation of model parameters that occurs before the acceptable model quality has been established.

59. (Currently amended) The method of claim 58, each iterative estimation until the acceptable model quality is established further comprises initializing the first training ~~algorithm~~ method by the same parameter values.

60. (Previously Presented) The method of claim 58, the controlling further comprises reusing at least some of the parameters obtained in a previous estimation of model parameters to initialize a subsequent estimation of model parameters for a next larger subset of the training data set.

61. (Canceled)

62. (Currently amended) A computer implemented method to facilitate constructing a statistical model, comprising:

separating computer readable data into a holdout data set and a training data set, the computer readable data is statistically characterizable;

iteratively estimating statistical model parameters for a subset of the training data set until a first convergence threshold is satisfied and evaluating the estimated statistical model parameters relative to the holdout data set;

repeating the estimation and evaluation of statistical model parameters obtained with successively larger subsets of the training data set until determining a size of data subset that provides acceptable statistical model parameters, acceptable statistical model parameters attained where the expected marginal cost outweighs the expected marginal benefit associated with successively larger subsets in order to facilitate reducing cost associated with clustering data relative to the computer readable data;

after determining the size of data subset that provides acceptable statistical model parameters, iteratively estimating statistical model parameters for a data subset of the acceptable size until a second convergence threshold is satisfied, the second convergence threshold being less than the first convergence threshold; and

based at least in part on the estimated statistical model parameters identified at the second convergence threshold, identifying a good clustering [[of]] data relative to the computer readable data to facilitate clustering data, wherein the clustered data is provided.

63. (Currently amended) A computer implemented system to facilitate building a statistical model for a computer readable data set, comprising:

first means for building a rough statistical model to characterize a subset of the computer readable data set;

~~evaluation~~ means for evaluating the acceptability of the rough statistical model based at least in part on an expectational cost-benefit analysis to facilitate reducing cost associated with clustering data relative to the computer readable data set, the first means building another rough statistical model for a larger subset of the data set if the evaluation means determines that a prior rough statistical model is unacceptable;

second means, which is different from the first means, for building a refined statistical model from an aggregate subset of data that yielded the rough statistical model deemed acceptable by the evaluation means; and

means for identifying a cluster of data within the computer readable data set based at least in part on the refined statistical model to facilitate clustering data relative to the computer readable data set, wherein the clustered data is provided.

64. (Currently amended) A computer implemented system to facilitate building a statistical model for a computer readable data set, comprising:

first means for estimating statistical model parameters from a subset of the computer readable data set, the data set is statistically characterizable;

means for evaluating the estimated statistical model parameters relative to a holdout data set of the data set;

means for determining a data subset from the training data set by causing the first means and the means for evaluating to respectively repeat estimation and evaluation of statistical model parameters with a next successively larger subset of the training data set until an acceptable quality of the statistical model parameters is established, the quality of the statistical model parameters established when the expected cost of generating the next successively larger subset outweighs the expected benefit in accuracy of utilizing the next successively larger subset <u>in order to facilitate reducing cost associated with clustering data relative to the computer readable data set</u>;

second means for estimating statistical model parameters based <u>at least in part</u> on the determined data subset to provide a more accurate estimation of model parameters than the first means;

means for setting parameters associated with cluster weights of a cluster of data; and[[;]]

means for determining the cluster of data contained in the computer readable data set based <u>at least in part</u> on the more accurate estimation of statistical model parameters <u>to facilitate clustering data relative to the computer readable data set, wherein the clustered data is provided</u>.